Privacy and Personal Data Collection with Information Externalities

Jay Pil Choi[†] Doh-Shin Jeon[‡] Byung-Cheol Kim[§]

January 29, 2019

Abstract

We provide a theoretical model of privacy in which data collection requires consumers' consent and consumers are fully aware of the consequences of such consent. Nonetheless, excessive collection of personal information arises in the monopoly market equilibrium which results in excessive loss of privacy compared to the social optimum. The main mechanism for this result is information externalities and users' coordination failure in which some users' decision to share their personal information may allow the data controller to infer more information about non-users. We also show that the emergence of data brokerage industry can facilitate the collection and monetization of users' personal data even in a fragmented market where no individual website has incentives to do so independently due to scale economies in data analytics. We discuss policy implications of our analysis in light of the recent EU General Data Protection Regulation (GDPR).

Key words: privacy, personal data, information externalities, GDPR

---

*We thank Jacques Crémer, Alexandre de Cornière, Saara Hämäläinen, Christian Peukert, David Laband, Marc Lebourges, Yassine Lefouili, Jean Tirole, Liad Wagman, and conference and seminar participants at the 2018 KEA-KAEA International conference, 2018 Mannheim Workshop "Governance of Platform Markets in the 'Big Data' Era", 2017 IIOC, 2017 IDEI-TSE-IAST conference on "The Economics of Intellectual Property, Software and the Internet," Academia Sinica, Auburn U, Korea U, U of Seoul, U of Central Florida, Toulouse School of Economics, and U of Alabama for helpful comments. We are also grateful to an anonymous referee and the Co-Editor Kai Konrad for constructive comments and guidance which greatly improved this article. Jay Pil Choi's research was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2016S1A5A2A01022389).

[†]Michigan State University, 220A Marshall-Adams Hall, East Lansing, MI 48824-1038 and School of Economics, Yonsei University, Seoul, Korea. E-mail: choijay@msu.edu.

[‡]Toulouse School of Economics, University of Toulouse Capitole, Manufacture de Tabacs, 21 allees de Brienne - 31000 Toulouse, France. E-mail: dohshin.jeon@gmail.com.

[§]Department of Economics, Finance, & Legal Studies, Culverhouse College of Business, University of Alabama. Tuscaloosa, AL, 35487, USA. E-mail: bkim34@cba.ua.edu.

1 Introduction

The Internet is now an essential component of our daily lives, and has profoundly changed the way we work, conduct our personal lives, and interact with other people. As we rely more on the Internet, it has become more of a necessity to have constant access to it via mobile devices and computers. However, one consequence of this development is that our routine online activities, such as email, search, and online shopping, constantly generate data about ourselves. The massive and unprecedented scale of personal data generation in conjunction with rapid reductions in computing costs for data storage and analytics naturally have led to serious privacy concerns by the public and policy-makers (Schneider, 2015).[1]

One puzzling aspect of this privacy debate is why people set aside their privacy concerns and voluntarily provide their personal information to websites and content providers despite their publicly stated objections and concerns about privacy loss (Singer et al. 2001; Waldo, Lin, and Millet 2007). Certainly there are often cases where data surveillance is taking place with neither our awareness nor consent, but it is also true that we frequently agree to it. For instance, we let Google have access to all metadata we generate in exchange for the use of Google applications such as Gmail, YouTube, Google Maps, etc. We also allow uninterrupted use of location tracking for most GPS-based services.[2]

Our study has been motivated by the following fundamental questions around these phenomena: When do firms collect too much personal data from a social planner's perspective? Why do people tend to allow some form of personal data collection which appears to harm themselves in the end? Do we expect that most individuals would no longer voluntarily agree to such data collection as soon as they become fully aware of the 'deals' each of them is making? Put differently, would it be enough to educate consumers about the exact costs of sharing their data for a socially desirable privacy protection? What are the appropriate policy remedies to address the privacy concerns from personal data collection and processing?

To address these questions, we develop a model of privacy with a monopolistic platform that sells content. Our model incorporates two-dimensional information heterogeneity to analyze the extent and types of information collected by the monopolist. First, there is a continuum of heterogeneous information types that differ in its sensitivity regarding pri-

[1]There are too many articles on privacy concerns published in the news media. For example, see Andrew Burt and Dan Geer, "The End of Privacy" (Oct. 5, 2017) *The New York Times*.

[2]People even allowed uninterrupted use of location and camera tracking to play once sensational augmented reality game Pokémon Go, which led to privacy concerns by regulators. For details, see the article by Sam Biddle, "Privacy Scandal Haunts Pokemon Go's CEO" (9 Aug 2016) *The Intercept*.

vacy cost. Second, we assume that there are two categories of information depending on whether the collection of information for some consumers enables the firm to infer information about others: information with externalities (E category) and information without externalities (N category). We show that the market equilibrium is characterized by too much data collection from the social planner's perspective, in particular, excessive collection of sensitive information in the E category. This equilibrium arises even if consumers are fully aware of the consequences of their consent. Our finding suggests that the provision of information or education would not fix the problem.

The primary mechanism for the results is information externalities. Some people's decision to share their personal information may allow the parties accessing to the information to know more or better about others, those who choose not to share their information (MacCarthy, 2011). Information externalities have been more potent due to significant advances in big data analytics which have made it possible to draw more accurate inference about those consumers who had not shared their data based on the data gleaned from those who had shared. In this environment, even if each user supposedly is aware of the potential harm of personal data release to herself, she may not take into account the entire spillover effects of her data release, either positive or negative, on other users. As a result, individually optimal decisions by even fully informed agents may not lead to a socially efficient outcome. If negative externalities exceed positive ones, the equilibrium will feature a situation where the data collection reaches socially harmful levels. If the opposite situation arises, by contrast, privacy concerns would deter socially desirable construction of greater data-networks.

We also consider an extension of the monopoly framework into an alternative market structure with a continuum of small websites in order to explore the role of data brokerage firms in the aggregation of information. In this set-up, we focus on how these information externalities operate at the level of entry of small websites. We show that even if each website alone has no incentives to collect personal data due to its small scale of operation, the emergence of data brokerage markets that purchase and aggregate data from multiple websites can restore incentives to collect personal data. The intuition behind this result is the same as in the monopoly model. The information externalities by each website lower each consumer's reservation utility evaluated when her data is not provided, which in turn reduces the compensation each website should make for consumer nuisance. As a result, even without any business stealing effects as in Mankiw and Whinston (1986), we can find an equilibrium in which too many websites enter to collect and sell personal data to data brokers.

Finally, we discuss policy implications of our study and effectiveness of policy remedies.

Specifically, we first show that a social planner's data policy regulation that requires consumers' explicit "opt-in" consent for the collection of sensitive information may not be an adequate remedy if the firm can offer a price discount to opt-in consumers. However, we also show that not allowing such price discrimination (that is, regulating the firm to offer the same price regardless a user's opt-in or the default choice) can alleviate the problem, but still may not completely recover the social optimum if the externality is strong enough. In such a case, the outright ban on data collection and trade on certain types of information can be a complementary policy lever to no discrimination regulation. We then connect our policy discussion to specific rules under the recent EU General Data Protection Regulation (GDPR).

Our research thus has important implications for the recent policy debate regarding data brokerage and privacy. The European Commission, for instance, introduced data protection policies which require websites to receive consumer approval for transferring personal data to third parties such as data brokerage firms. The U.S. Federal Communications Commission passed a similar rule. The new rule requires Internet providers such as AT&T, Verizon, and Comcast to obtain their customers' explicit consent before using or sharing sensitive data with third parties such as marketing firms, which was one of their main sources of revenue generated by turning customers' behavioral data into better information basis for targeted advertising.[3] These policies may have some effects on naive consumers by alerting them to be aware of such data transfers. Nonetheless, the overall effects of such a consent-based approach may be limited in addressing the negative information externalities problem since well-informed, fully rational consumers may not change their behaviors because opting-out may not be individually rational in the presence of information externalities.

As in our paper, several legal scholars pointed out the public good nature of privacy and warned of the ineffectiveness of the 'informed consent model' as a solution to protect against invasion of privacy.[4] In essence, this notice-and-consent approach is based on the premise that each individual should have control for disclosure and dissemination of his own personal information. However, this individualistic choice approach is inadequate in addressing entire privacy concerns due to information externalities. MacCarthy (2011),

[3]For the EU policy, see Directive 95/46/EC (the data protection Directive). For the FCC's new privacy ruling, see Brian Fung and Craig Timberg, "The FCC just passed sweeping new rules to protect your online privacy." (27 Oct. 2016) *The Washington Post*.

[4]It appears that there is no universally accepted terminology: 'notice-and-choice' and 'notice-and-consent' approach are other terms often used interchangeably. Essentially, they all describe the same approach that data operators should inform individuals of the data policy and each individual decides to agree to it or not.

for instance, argues that the reliance on individual consent to determine the collection and use of personal information will be ineffective in the presence of negative information externalities and potential risks of information leakage. In a similar vein, Fairfield and Engel (2015) propose to label privacy as a public good and thus call for a collective choice approach to address the privacy issue. Earlier, Hirsch (2006) also pointed out similarities between privacy regulation and environmental laws. Our paper formalizes these ideas.

1.1 Related Literature

The literature on privacy is vast and extensive. We thus do not intend to provide an exhaustive review of the literature. Instead, we limit our discussion to selective strands of literature more directly related to this article. For more comprehensive reviews of economic perspectives on privacy and the Internet, we refer to Athey (2014) and Acquisti, Taylor, and Wagman (2016).[5]

One branch of literature to which we make a meaningful contribution is the recent debate on how we should address privacy concerns against prevailing data broker industry. According to the U.S. Senate (2013) and the White House (2014), "data brokers" have vibrantly collected, packaged, and traded sensitive consumer data mostly behind a veil of secrecy and thus pose substantial privacy concerns for consumers. The supply chain appears to start from the interactions between web users and web-based applications/content providers of which business model consists in monetizing personal digital trails. Those websites feed the collected data to data brokers who then sell the data after some processing to interested third parties such as advertisers and marketers. The use of those data is not expected to be limited to designated purposes only, and there could be further transfers to others. In this article, we focus on the very early stage when each user voluntarily agrees to uncommitted data use; we aim to provide an economic rationale for the users' consent to the websites.[6]

Certainly, there should be some convincing behavioral reasons why users give away their personal data. Some have mentioned consumers' lack of understanding about websites' data use policy. For instance, the Australian Information Commissioner and Privacy Commissioner Timothy Pilgrim remarked that privacy notices are just too long for people to read through and most people find it difficult to understand what they are signing up to.[7]

[5]For reviews with broader perspectives, we find Lane et al. (2014) and Smith, Dinev, and Xu (2011) very helpful. For behavioral approaches to privacy issues, see Acquisti (2009) and Acquisti and Grossklasgs (2004, 2007).

[6]Voluntary over-disclosure in web-forms was also found in a field experiment (Preibusch, Krol, Beresford, 2013).

[7]"Many privacy policies are long, complex: OAIC." *ZDNET*. (Aug. 15, 2013) by Corinne Reichert. http://www.zdnet.com/article/many-privacy-policies-are-long-complex-oaic/

Alternatively, it could be due to consumers' myopic and time-inconsistent preference. For discussion's sake let us envision a user's data sharing decision from the perspective of 'contract design and self-control' à la Dellavigna and Malmendier (2004). Then, users would view the enticing free (or highly subsidized) content services as 'leisure goods' that provide immediate benefits but impose delayed costs of privacy loss. Any naive time-inconsistent users will easily opt for enjoying the free content services now by agreeing to the data use policy even if they are well aware of the future costs. Another explanation put forward is that the costs of privacy loss are at best nebulous and intangible so that the users end up underestimating them substantially. Admitting the persuasiveness of those behavioral expositions, for the purpose of this paper we rather assume away any bounded rationality or consumers' lack of knowledge about the website's data use. In other words, we assume the very rational consumers who are fully aware of all consequences from their choices so that there is no way to resort to consumers' myopia or to limited information. Even so, we show that each consumer can find it individually rational to accept the third party use of their personal data and that in general there is socially excessive monetizing of personal digital data in the presence of negative information externalities. Thus, we provide a theoretical foundation calling for a different policy beyond the current notice-and-choice approach, which takes the same stance as Hirsch (2006), MacCarthy (2011) and Fairfield and Engel (2015).[8]

Our research is also associated with the literature on data acquisition and pricing which mostly adopts two-sided market configurations. For instance, Bergemann and Bonatti (2015) consider a model of data provision and data pricing in which a single data provider controls a large database about the match value information between individual consumers and individual firms. They analyze the equilibrium data acquisition and pricing policies when such information allows targeted advertising. Their focus is on the data provider's optimal pricing policy and how the price of data influences the composition of the targeted set, but do not address the issue of privacy. Since we focus on information externalities on the consumer side and how the database can be aggregated through the data brokerage markets, our work is complementary to theirs, for the two works combined provide a more comprehensive perspective on the use of consumer data. Our work also reminds of Bataineh et al. (2016) in that they propose a data monetization platform intermediating individuals (personal data sellers) with merchants (data users). We explore the adverse effects of data monetization on individual privacy whereas they view active data trading as a

[8]Campbell, Goldfarb, and Tucker (2015) suggest a new distortion by the commonly used consent-based approach that may disproportionately benefit big firms but adversely affect small and new firms.

potential market mechanism for higher profits for both data sellers and buyers.[9]

Our work is related to recent studies where firms benefit from better targeting but consumers try to avoid privacy costs. For examples, Goh, Hui, and Png (2016) empirically examine the effects of privacy and marketing externalities from the U.S. Do Not Call registry. Johnson (2013) studies the interplay of targeted advertising by merchants and advertising avoidance by consumers who can install ad-blockers. Montes, Sand-Zantman, and Valletti (2015) study a Hotelling-type duopoly model where competing firms can acquire information about consumers' characteristics for a better personalized pricing while consumers can pay a 'privacy cost' to avoid such price discrimination. As these studies have in common, consumers may take various actions to avoid the costs of privacy loss. In our model, in the absence of information externalities, each consumer can avoid all privacy costs by choosing no consumption, but with heterogeneous valuation on consumption, the high valuation consumers decide to sacrifice the privacy for entertaining the content service. In our model, the consumers are rather passive in the sense that their choice is limited and do not take active actions considered in these papers. It seems an interesting future research to study the interplay between information externalities, ads avoidance, and targeted marketing.

We also notice that many economists studied various privacy issues in the Internet such as the effects of competition on privacy (Casadesus-Masanell and Hervas-Drane 2015), impact of taxation on data collection (Bloch and Demange 2018; Bourreau, Caillaud, and De Nijs 2018), effects of a privacy regulation on firm's investment in quality (Lefouili and Toh 2017). Generally our research adds to this burgeoning literature on information, privacy, and the Internet.

The rest of the paper is organized as follows. We discuss the information externalities in Section 2. Then we introduce the model of a monopolist in Section 3. We analyze the comparison between social planner's and the monopoly firm's optimal data collection policy concerning consumer types and information types in Sections 4 and 5, respectively. In Section 6, we briefly review the main finding from the data brokerage model of which the analysis is presented in the Appendix. In Section 7 we study policy remedies and discuss policy implications related to GDPR. Section 8 contains concluding remarks.

2 Information Externalities

In the model that we will introduce in the next section, we assume that consumers incur a nuisance cost of privacy loss when personal information is collected and used. There can be

[9] In our model, consumers can get subsidized for their web-content when they agree to the uncommitted data use policy, but they do not actively seek monetary compensation by selling their personal data as a valuable economic good via the intermediating platform enabling the data aggregation for a higher valuation.

many sources of such utility loss. For instance, there could be direct economic losses due to personalized pricing enabled by the detailed knowledge of personal preferences. This kind of loss reminds of the classic argument by the Chicago School scholars (e.g., Posner (1978, 1981) and Stigler (1980)) that one main reason for demanding privacy is to avoid exploitation by potential trading partners who might take advantage of the released information against the revealing individuals.[10] We can also think of a variety of psychological reasons for negative feelings about privacy loss. A newly released smartphone app called Google Trips, promoted to provide a "personalized tour guide in your pocket," is a case in point. The modus operandi of this app developed by Google is to "use what it already knows about you, based on data it has collected from your Gmail account, and combines it with established features from its other offerings, like Destinations, and its large database of crowd-sourced reviews," which led a New York Times reviewer for this app to comment that "It's Kind of Creepy."[11]

The key driving force in our paper is information externalities which are often referred to as 'negative privacy externalities' in the literature of data breach and privacy concerns (MacCarthy, 2011). More specifically, this concept is based on the premise that some people's decisions to share their personal information may allow the data controller to know more or better about others whose direct data is not obtained yet. The externality arises because (dis)utility of those who did not share their personal data can be affected by those who shared the data.[12]

One example illustrating such a mechanism is a study by MIT students who showed that men's sexual orientation can be predicted by an analysis of social network sites such as Facebook. This is possible because data analytics reveal that homosexual men have proportionally more gay friends than straight men, which allows one to predict men's sexual orientation based solely on the sexuality of their friends (Jernigan and Mistree, 2009). In fact, Hal Abelson, a computer science professor at MIT, responded to this study by saying

[10]Privacy protection in this context is likened to the protection of fraudulent claims by Posner as there are no social benefits from privacy protection (except a taste for privacy itself). In contrast to this disclosure literature, there is a vast literature on costly screening and distortionary signaling about private types. Daugherty and Reingaum (2010) provide a new model of the economics of privacy related to both strands of literature.

[11]To quote, "Before you create your first trip, you'll see some of your previous trips that you didn't even share. That's because it has already pulled in information from your Gmail account, so it knows which hotels you stayed in and where you rented a car from and stores this information under Reservations." See Justin Sablich, "How to Use Google to Plan Your Trip," (21 Sep 2016) *New York Times*.

[12]Information externalities are similar to data spillovers in Tucker (2017). She considers a scenario in which a person takes a picture of her car with geocode using an app after parking to help her remember the exact parking spot. However, the photo may record other people and cars, and they may be identifiable through facial recognition or license plate databases, creating potential spillovers for others who did not take the photo. As a result, their privacy may be compromised.

that "you don't have control over your information" [13] even though you do not divulge your personal information, if other people do. This interpretation features exactly the negative information externalities in our study. In a similar vein, Kosinski, Stillwell, and Graepel (2013) show that 'Facebook Likes' can be used to accurately predict a range of highly sensitive personal attributes such as sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.[14]

Genetic tests are another example of privacy concerns due to informational externalities. Researchers have found that some subjects' genetic information can be used to make predictions of others' genetic disposition among the same racial or ethnic category. Erlich et al. (2018), for instance, show that only two percent of the population needs to have done a DNA test to identify nearly everyone else. Because of practical concerns about privacy and/or invidious discrimination based on genetic information, the U.S. federal government has prohibited insurance companies and employers from any misuse of information from genetic tests under the Genetic Information Nondiscrimination Act (GINA).

The standard adverse selection mechanism can also operate through our information externality channel. A device called 'telematics' can track detailed driving habits of car insurance customers. Some drivers would agree to install the tracking device for a premium discount. This decision, however, can affect not only the installing driver's insurance premium but also the other drivers'. This is because normally more careful drivers would adopt the installation of a telematics device, which implies that no adopting drivers are considered not as much as careful and face a higher premium. Progressive's Snapshot program and Allstate's Drivewise Mobile App are the real-world examples of this kind. Reimers and Shiller (2018) provides many examples of telematics in auto insurance.

"Doppelganger search" is an example of data analytic techniques that can create information externalities. Consider the situation that firms may have from the beginning some information about consumer i, which they obtained from data available from off-line or online using public or private sources. Then, they can find some consumer i' whose personal data matches consumer i (up to the information they have about consumer i). Hence, even if consumer i does not use the service, the fact that there are several consumers similar to i who use the service may allow some inference about consumer i. This process is feasible by a zooming-in with big data with artificial intelligence algorithm (Stephens-Davidowitz,

[13] Johnson, Caroly Y. "Project 'Gaydar'" (20 Sep 2009) *The Boston Globe*.

[14] Once Scott McNealy, a co-founder of Sun Microsystems, even uttered plainly back in 1999 that "You have zero privacy anyway. Get over it." See the article by Polly Sprenger (26 Jan 1999), "Sun on privacy: 'Get over it,' " *Wired*.

2017).[15]

3 The Monopoly Model

We consider a monopolistic online platform offering content service. There is a mass one of consumers whose valuations for the service, denoted by u, are distributed over $[\underline{u},\overline{u}]$ with distribution function F and density f. We assume that F satisfies the standard monotone hazard rate condition (MHRC), that is, $f/(1-F)$ is non-decreasing. The monopolistic content provider can collect consumers' personal data in the process of providing its service, which then can be potentially utilized for other purposes such as targeted advertising and promotion of other ancillary services. Let m measure the mass of consumers who use the service, that is, $m = 1-F(u)$ where u is the cutoff type of consumers such that all consumers whose valuation exceeds or is equal to u use the service. If $\underline{u}$ is sufficiently high, then the market is fully covered ($m = 1$).

Let Θ represent the set of the types of personal information that the platform can gather from each user. We rank different types of information according to θ, which represents the sensitivity in terms of privacy cost. We assume that θ is distributed over $[0,1]$ with distribution function G and density g. Suppose that the platform collects the information type θ from all its users of mass m. Then, each user suffers from nuisance costs $\psi(\theta,m) > 0$ either from privacy concerns such as data breach accidents or privacy costs originating from the user's per se taste for privacy. We assume that $\psi(\theta,m)$ is a strictly increasing function of θ. The dependence of $\psi(\cdot)$ on m is motivated by information externalities among users generated by data analytics and we assume that $\psi(\cdot)$ is increasing in m. To allow the possibility that personal data collected by the firm can be used to provide better service to consumers, we assume that consumers can obtain benefit $b(\theta,m) \geq 0$. We do not make any particular assumption about the dependence of b on θ because collection of more sensitive information (such as social security number) does not necessarily lead to better service. For instance, more basic information (such as gender and age group) may be more useful in providing better tailored service for consumers. Even when $b(\theta,m)$ is increasing in θ, we assume that nuisance costs are more responsive to θ, that is, the net nuisance, $\lambda(\theta,m) = \psi(\theta,m) - b(\theta,m)$, is a strictly increasing function of θ, and that $\lambda(\cdot)$ increases with m. We assume that the data collection is beneficial for $\theta = 0$, i.e., $\lambda(0,m) < 0$, but the

[15]The business practice of "steering" (U.S. Council of Economic Advisers, 2015) can be viewed as another example of information externalities with the application of doppelganger search. Zarsky (2018) outlines how data profiling can generate information externalities. Suppose that a customer visits a website, shares personal demographic information and zip code, and buys high-priced, low-quality products. Then, another customer who has a similar demographic and zip code is steered to see high-priced, low-quality products first at the same website.

net nuisance becomes positive and sufficiently large as θ increases. This assumption means that data with low θ are essential for consumer benefit while data with high θ generate large net nuisances.

One of the main drivers in our model is the information externality effect of data collection on *non-users* of the service. To analyze this effect, we incorporate another dimension of data heterogeneity in the extent of informational externalities on non-users. For each type of θ information, there are two categories of data: α proportion of each θ type information generates information externalities on non-users, whereas $(1-\alpha)$ proportion does not. For example, genetic data, mobile use or payment data, surveys on political views are more likely to generate externalities than others. As it will be clear below, the monopolist's and social planner's optimal data collection policies will be characterized by cut-off values of θ for each category of data. Let θ_E and θ_N denote the cutoff values of θ for externality (E) and no-externality (N) category data where the information type $\theta \in [0,\theta_E]$ in externality category and $\theta \in [0,\theta_N]$ in no-externality category are collected for the use of the service. Then, the measure of information collected can be written as

$$h = \alpha \int_0^{\theta_E} dG + (1-\alpha) \int_0^{\theta_N} dG = \alpha G(\theta_E) + (1-\alpha) G(\theta_N). \tag{1}$$

A service user's nuisance costs are assumed to be additive and given by

$$\Psi(\theta_E, \theta_N, m) = \alpha \int_0^{\theta_E} \psi(\theta, m) dG + (1-\alpha) \int_0^{\theta_N} \psi(\theta, m) dG. \tag{2}$$

Thus, $\Psi(\theta_E, \theta_N, m)$ captures each user's nuisance costs when the mass of total users is m, and each category of data is collected with cut-off values of (θ_E, θ_N). Similarly, the total benefit of a user from data collection can be written as

$$B(\theta_E, \theta_N, m) = \alpha \int_0^{\theta_E} b(\theta, m) dG + (1-\alpha) \int_0^{\theta_N} b(\theta, m) dG.$$

The total net nuisance costs are given by

$$\begin{aligned} \Lambda(\theta_E, \theta_N, m) &= \alpha \int_0^{\theta_E} \lambda(\theta, m) dG + (1-\alpha) \int_0^{\theta_N} \lambda(\theta, m) dG \\ &= \Psi(\theta_E, \theta_N, m) - B(\theta_E, \theta_N, m) \end{aligned}$$

Due to information externalities, non-users of the service can also be subject to nuisance or privacy costs. Let $\widehat{\psi}(\theta, m) > 0$ denote a non-user's nuisance costs from the collection

of type θ in the externality data category.[16] For simplicity, we assume that

$$\widehat{\psi}(\theta, m) = \xi \psi(\theta, m), \tag{3}$$

where $\xi \in (0,1)$ represents the extent of information externalities on non-users. With additive nuisance costs, a non-user's nuisance costs can be written as

$$\widehat{\Psi}(\theta_E, m) = \alpha\xi \int_0^{\theta_E} \psi(\theta, m) dG. \tag{4}$$

We assume that the monopolist can generate additional revenue of $R(h,m)$ when the personal data is utilized, with $\frac{\partial R}{\partial h} > 0$ and $\frac{\partial R}{\partial m} > 0$. That is, we assume that each type of information contributes to the monopolist's revenue equally, and thus the monopolist's revenue only depends on the aggregate amount of information h collected per user and the measure of users m. We also assume that $R(h,m)$ is concave in each element.[17] We further assume that $R(h,m)$ represents social benefits as well. Alternatively, we could assume that there exist other channels through which social benefits are generated from the collected data. By assuming away such consideration, we can focus on the consumers' coordination failure and negative externalities in the nuisance costs.[18] As we are mainly concerned with a digital product/service, the marginal cost of the content is assumed to be zero.

The monopolist makes choices on both data collection policy and monopoly pricing. Given its data collection policy, the monopolist chooses the subscription price p, which determines the measure of consumers who use the service (m) as in the standard monopoly problem. In addition, given m, it chooses its data collection policy by choosing the vector of cutoffs $\theta = (\theta_E, \theta_N)$. In what follows, we analyze the monopolist's choices and compare them to the socially optimal ones.

4 Data Collection Policy concerning Consumer Types

In this section, we study the optimal pricing for a *given* data collection policy.

[16] In principle, we could also introduce positive externalities on non-users for some type of information. We briefly discuss this extension in the remark provided at the end of Section 5.

[17] As the reviewer pointed out, $R(h,m)$ may depend on the market structure of the product in question. For instance, if the product market is characterized by symmetric oligopoly competition and consumer-specific information is used for price discrimination, competition may be intensified compared with uniform pricing, as in Thisse and Vives (1988). We sidestep this issue by assuming a monopolistic market structure. In addition, we consider a situation in which the additional revenue source represented by $R(h,m)$ is through advertising while maintaining the assumption of uniform pricing.

[18] With additional benefits generated from the collected data, the socially desirable level of data collection would increase compared to the currently assumed situation. This consideration would not affect qualitative results of this paper.

4.1 Social Optimum: The First Best Benchmark

We first analyze the socially optimal outcome as a benchmark in which a social planner chooses the allocation (i.e., the set of consumers who use the service), given a data policy characterized by $\theta = (\theta_E, \theta_N)$. The social planner chooses the cutoff type of consumers, u, such that all consumers whose valuation exceeds or is equal to u use the service, i.e., $m = 1 - F(u)$. Social welfare given a cutoff type u is given by

$$\begin{aligned} W(u,\theta) &= \int_u^{\overline{u}} x dF(x) + R(h, 1-F(u)) \\ &\quad - (1-F(u))\,\Lambda(\theta, 1-F(u)) - F(u)\widehat{\Psi}(\theta_E, 1-F(u)). \end{aligned}$$

The first term captures the aggregate total intrinsic utility for all users. The second term is the firm's revenue from data collection of measure h from m mass of consumers. The third term captures the total net nuisance costs of users, and the last term is the total nuisance costs of non-users.

The welfare-maximizing cutoff type u^s can be derived by the first order condition with respect to u, which yields

$$u + \frac{\partial R}{\partial m} = \underbrace{(\Lambda - \widehat{\Psi}) + (1-F(u))\frac{\partial \Lambda}{\partial m} + F(u)\frac{\partial \widehat{\Psi}}{\partial m}}_{\text{social marginal cost } (SMC_m) \text{ of net nuisance from additional consumer}} . \tag{5}$$

The RHS of (5) represents the social marginal cost (SMC_m) of net nuisance when one additional user joins the customer base. There are three channels through which SMC_m is affected when an additional user joins to use the content service. First, the marginal consumer's status change from a non-user to a user directly affects his net nuisance cost by $(\Lambda - \widehat{\Psi})$. In addition, a new user inflicts externalities not only on the user group he joins, but also on the non-user group he leaves behind. The net nuisance cost of an existing user changes by $\frac{\partial \Lambda}{\partial m}$ as a new user joins, with the aggregate change for the user group being equal to $(1-F(u))\frac{\partial \Lambda}{\partial m}$. The nuisance cost of an existing non-user also changes by $\frac{\partial \widehat{\Psi}}{\partial m}$ with the aggregate effect being $F(u)\frac{\partial \widehat{\Psi}}{\partial m}$. The last two terms represent these effects of negative information externalities.

4.2 Monopolist's Optimal Choice

For a given data collection policy of $\theta = (\theta_E, \theta_N)$, we now consider the monopolist's optimal choice of the content price which will determine the number of users. Given the monopolist's content price p, let u be the cutoff type of consumer who is indifferent between using the service or not. For the cutoff utility u, the individual rationality (IR) constraint

can be written as

$$u - p - \Lambda(\theta, 1 - F(u)) \geq -\widehat{\Psi}(\theta_E, 1 - F(u)),$$

where $-\widehat{\Psi}(\theta_E, 1 - F(u))$ is the reservation utility of type u consumer. As the IR constraint is binding, the price is given by

$$p = u - \Lambda(\theta, 1 - F(u)) + \widehat{\Psi}(\theta_E, 1 - F(u)). \tag{6}$$

The monopolist solves the following problem:

$$\underset{u}{Max}\, \Pi(u, \theta) = (1 - F(u))p + R(h, 1 - F(u))$$

where p is defined in (6). The first order condition for profit maximization is given by:

$$\begin{aligned} \frac{\partial \Pi(u,\theta)}{\partial u} &= -f(u)(u - \Lambda + \widehat{\Psi}) \\ &+ (1 - F(u)) \left(1 + \frac{\partial \Lambda}{\partial m} f(u) - \frac{\partial \widehat{\Psi}}{\partial m} f(u) \right) - \frac{\partial R}{\partial m} f(u) = 0. \end{aligned}$$

Define $u - \frac{1-F(u)}{f(u)} \equiv u^v(u)$ to be the "virtual valuation" of a consumer with value u. Then, the profit-maximizing cut-off type u^* can be obtained by rearranging the first order condition as follows:

$$\underbrace{u^v(u)}_{\text{virtual valuation}} + \frac{\partial R}{\partial m} = \underbrace{(\Lambda - \widehat{\Psi}) + (1 - F(u)) \left(\frac{\partial \Lambda}{\partial m} - \frac{\partial \widehat{\Psi}}{\partial m} \right)}_{\text{private marginal cost } (PMC_m) \text{ of net nuisance from additional consumer}} . \tag{7}$$

Let $m^* = 1 - F(u^*)$. Note that if we consider the standard monopoly model without additional source of revenue from personal data use and nuisance costs, that is, $b(\cdot) = \psi(\cdot) = \widehat{\psi}(\cdot) = R(\cdot) = 0$, condition (7) reduces to the standard monopoly condition, $u^v(u) = 0$.[19] In contrast, for the monopolist in our model, the LHS of (7) becomes $u^v(u) + R'$ to reflect the additional revenue R' from data monetization. The RHS represents the private marginal cost (PMC_m) of net nuisance from data collection of additional consumer. The comparison of (5) and (7) shows a new type of distortion that makes the private marginal

[19] Because the virtual valuation is non-decreasing with u under MHRC, there is a unique solution to the monopoly problem.

cost differ from the social one.

$$SMC_m - PMC_m = F(u)\frac{\partial \widehat{\Psi}}{\partial m} + [1 - F(u)]\frac{\partial \widehat{\Psi}}{\partial m} = \frac{\partial \widehat{\Psi}}{\partial m} > 0. \tag{8}$$

When one extra consumer is served and his data is added to the monopolist's database, it inflicts additional negative externalities to $F(u)$ measure of non-users even though they do not use the monopolist's content. This effect on non-users' reservation utility is $F(u)\frac{\partial \widehat{\Psi}}{\partial m}$. While the social planner cares about these negative externalities, the monopolist does not because they are non-users. Instead, the monopolist cares about the effect of an additional user on its ability to extract surplus from existing users. However, this is no concern to the social planner because it is just a pure transfer. In order to induce each existing consumer to keep consuming the content, the monopolist's price needs to be adjusted below by $(\frac{\partial \Lambda}{\partial m} - \frac{\partial \widehat{\Psi}}{\partial m})$ to compensate the differences in the nuisance cost change. Note that as the additional user also negatively affects non-users and reduces the reservation value of the marginal consumer, the price compensation needs to be only $(\frac{\partial \Lambda}{\partial m} - \frac{\partial \widehat{\Psi}}{\partial m})$, not $\frac{\partial \Lambda}{\partial m}$. As a result, the negative profit impact via a reduced price to the user group is given by $(1-F(u))(\frac{\partial \Lambda}{\partial m} - \frac{\partial \widehat{\Psi}}{\partial m})$ whereas the social planner only cares about the real impact on the user group which is $(1-F(u))\frac{\partial \Lambda}{\partial m}$. This creates an additional difference of $(1-F(u))\frac{\partial \widehat{\Psi}}{\partial m}$.

Taken together, the total difference between SMC_m and PMC_m becomes $F(u)\frac{\partial \widehat{\Psi}}{\partial m} + (1 - F(u))\frac{\partial \widehat{\Psi}}{\partial m} = \frac{\partial \widehat{\Psi}}{\partial m}$. Thus, this distortion pushes the monopolist to serve too many consumers and the extent to which the monopolist's decision departs from the social planner's depends on the additional user's impact on the reservation utility. The effect of this distortion is in the opposite direction of the standard monopoly result that the monopolist serves too few consumers. Let $m^s = 1 - F(u^s)$ and $m^* = 1 - F(u^*)$. We have:

Proposition 1 *In a standard monopoly model without additional revenue from data collection, the monopolist serves too few consumers relative to the social optimum. However, in the presence of the additional revenue from data collection, the monopolist can serve too many consumers and thus collect excessive personal data compared to the socially efficient level i.e., $u^* < u^s$ (equivalently, $m^* > m^s$) if information externalities are large enough.*

To illustrate the result of Proposition 1, consider a following simple parametric example:

Example 1. Let us assume that $\alpha = 1$, that is, all personal data belong to the externality category. Further assume that $R(h,m) = rhm$, where $r > 0$ is a revenue parameter, $u \sim \mathscr{U}[0,1]$, $\theta \sim \mathscr{U}[0,1]$, $\psi(\theta,m) = \eta\theta m$, $\widehat{\psi}(\theta,m) = \xi\psi(\theta,m) = \xi\eta\theta m$, where $\eta > 0$ is a nuisance cost parameter. With only one category of data with externality, let θ be the cut-off type data. Then, $h = \theta$. Assume $b(\theta,m) = b$ where $b > 0$ is a very small constant. We

thus have $B(h,m) = bh$. In this simplified version, we can derive the nuisance costs:

$$\begin{aligned}\Psi(\theta,m) &= \int_0^\theta \psi(x,m)dG = \frac{\eta\theta^2}{2}m \\ \widehat{\Psi}(\theta,m) &= \int_0^\theta \widehat{\psi}(x,m)dG = \xi\Psi(\theta,m) = \xi\left(\frac{\eta\theta^2}{2}m\right).\end{aligned}$$

As the difference between SMC_m and PMC_m is generated by $\widehat{\Psi}(\theta,m)$ only and does not depend on $B(h,m)$, this model is equivalent to the one in which $R(h,m) = (b+r)hm$ and $B(h,m) = 0$. In what follows, we consider the latter model when we compute *SMC* and *PMC*. Then, we have

$$\begin{aligned}SMC_m &= (1-\xi)\eta m\theta^2 + \xi\eta\frac{\theta^2}{2}, \\ PMC_m &= (1-\xi)\eta m\theta^2,\end{aligned}$$

with $SMC_m - PMC_m = \partial\widehat{\Psi}(\theta,m)/\partial m = \xi\eta\frac{\theta^2}{2}$.

The socially optimal outcome m^s (for a given θ) satisfies

$$(1-m^s) + (r+b)\theta = (1-\xi)\eta m^s\theta^2 + \xi\eta\frac{\theta^2}{2}. \tag{9}$$

The monopolist's outcome m^* (for a given θ) satisfies

$$(1-2m^*) + (r+b)\theta = (1-\xi)\eta m^*\theta^2. \tag{10}$$

By comparing the two first order conditions for the socially optimal outcome and the monopolist's choice, we can easily verify that we have $m^*(\theta) > m^s(\theta)$ if the following condition holds:

$$SMC_m - PMC_m = \xi\eta\frac{\theta^2}{2} > \frac{1+(r+b)\theta}{2+(1-\xi)\eta\theta^2}. \tag{11}$$

This condition can be satisfied for any $\xi \in (0,1]$ and $\theta \in (0,1]$ if η is sufficiently large. In (11), the LHS is linearly increasing in ξ whereas the RHS is increasing and convex in ξ. And at $\xi = 0$, the RHS is larger than the LHS. Therefore, if $\eta\theta^2 > 1+(r+b)\theta$ holds, there exists a unique $\xi^*(\theta) \in (0,1)$ such that $m^*(\theta) > m^s(\theta)$ if and only if $\xi > \xi^*(\theta)$.

This example indicates that the monopolist may serve too many consumers if the extent of informational externalities (ξ) and the nuisance cost parameter (η) are sufficiently large to outweigh the standard monopoly distortion effect.

5 Data Collection Policy concerning Information Types

In this section, we study the choice of the data collection policy concerning information types given a measure of users.

5.1 Socially Optimum Data Collection: The First-Best Benchmark

Given the measure of users (m), the welfare-maximizing cutoff types for each category of data, $\theta^s = (\theta_E^s, \theta_N^s)$, can be characterized by the following first order conditions:

$$\begin{aligned}
\frac{\partial W(u,\theta)}{\partial \theta_E} &= \left[\frac{\partial R}{\partial h} - m\lambda(\theta_E, m) - (1-m)\widehat{\psi}(\theta_E, m)\right] \alpha g(\theta_E) = 0, \\
\frac{\partial W(u,\theta)}{\partial \theta_N} &= \left[\frac{\partial R}{\partial h} - m\lambda(\theta_E, m)\right] (1-\alpha) g(\theta_N) = 0.
\end{aligned}$$

With $\widehat{\psi}(\theta_E, m) = \xi\,\psi(\theta_E, m)$, these two first-order conditions can be simplified as

$$\frac{\partial R}{\partial h} = m\lambda(\theta_E, m) + (1-m)\xi\,\psi(\theta_E, m), \tag{12}$$

$$\frac{\partial R}{\partial h} = m\lambda(\theta_N, m). \tag{13}$$

We define the social marginal cost of net nuisance for each category of data as follows:

$$SMC_E = m\lambda(\theta_E, m) + (1-m)\xi\,\psi(\theta_E, m), \tag{14}$$

$$SMC_N = m\lambda(\theta_N, m). \tag{15}$$

Assuming that R is concave in h, the above conditions show that $\theta^s = (\theta_E^s, \theta_N^s)$ are uniquely determined (as we assume that ψ and $\lambda = \psi - b$ are increasing in θ). From the social welfare perspective, as the marginal benefit of additional information collection is the same, the comparison boils down to the degree of information externalities generated from each data category. Because the collection of the externality category data inflicts negative externalities on the non-user group, the social planner will choose $\theta_E^s < \theta_N^s$ for $m \in (0,1)$. Under full market coverage (i.e., $m = 1$), however, there are no non-users subject to information externalities. As a result, we have $SMC_E = SMC_N$, which leads to $\theta_E^s = \theta_N^s$. We thus have the following results which characterize the social preference on the data collection policy.

Proposition 2 *A social planner will collect less of the data with negative externalities in comparison to the data with no externalities, i.e., $\theta_E^s < \theta_N^s$ for any $m \in (0,1)$. Only when the market is fully covered with $m = 1$, the social planner will collect the same amount of data for both categories with $\theta_E^s = \theta_N^s$.*

The amount of data collection associated with the choices of θ_E^s and θ_N^s is calculated from (1):

$$h^s = \alpha G(\theta_E^s) + (1-\alpha)G(\theta_N^s). \tag{16}$$

5.2 Monopolist's Optimal Choice

Recall that the monopolist's total profit is given by

$$\Pi(\theta, m) = R(h, m) + mp$$

where $m = 1 - F(u)$, $p = u\ -\Lambda(\theta, 1-F(u)) + \widehat{\Psi}(\theta_E, 1-F(u))$, and $h = \alpha G(\theta_E) + (1-\alpha)G(\theta_N)$.

The profit-maximizing cutoff types θ_E^* and θ_N^* can be characterized from the first order conditions. First, regarding θ_N^*, the F.O.C. is given by:

$$\begin{aligned} \frac{\partial \Pi(\theta, m)}{\partial \theta_N} &= \frac{\partial R}{\partial h}\frac{\partial h}{\partial \theta_N} + m\frac{\partial p}{\partial \theta_N} \\ &= \left[\frac{\partial R}{\partial h} - m\lambda(\theta_N, m)\right](1-\alpha)g(\theta_N) = 0, \end{aligned}$$

which yields

$$\frac{\partial R}{\partial h} = m\lambda(\theta_N, m). \tag{17}$$

This condition shows that the private marginal cost of net nuisance is the same as the social one for no externality (N) category data as is derived in (14):

$$SMC_N = PMC_N$$

Hence, we have $\theta_N^* = \theta_N^s$ with all other things being equal (i.e., if $m^* = m^s$ and $\theta_E^* = \theta_E^s$). For non-externality category data, an increase in θ_N has no impact on the reservation utility of consumers. In the absence of any informational externalities, there is no difference between the social optimum and the private choice.

Regarding θ_E^*, we can derive the following first-order condition:

$$\frac{\partial \Pi(\theta, m)}{\partial \theta_E} = \left[\frac{\partial R}{\partial h} - m\left[\lambda(\theta_E, m) - \xi\,\psi(\theta_E, m)\right]\right]\alpha g(\theta_E) = 0,$$

which yields

$$\frac{\partial R}{\partial h} = m\left[\lambda(\theta_E, m) - \xi\,\psi(\theta_E, m)\right]. \tag{18}$$

The LHS of (18) is additional revenue from collecting more data with an increase in θ_E.

The RHS of (18) captures the private marginal cost of net nuisance for the E category data, which is defined as:

$$PMC_E = m\left[\lambda(\theta_E, m) - \xi\,\psi(\theta_E, m)\right].$$

The comparison between (17) and (18) implies that the monopolist will collect more data in E category relative to data in N. In addition, the comparison between SMC_E from (15) and PMC_E shows that the social marginal cost of collecting data with informational externalties exceeds the private one, $PMC_E < SMC_E$. Hence, we can derive the following relationship:

$$PMC_E < PMC_N = SMC_N < SMC_E \tag{19}$$

The relationship captured in (19) clearly shows the role of the information externalities from the E category data: (i) they make SMC_E larger than SMC_N, (ii) they make PMC_E smaller than PMC_N because the information externalities worsen each non-user's reservation utility which makes the monopolist need to make less than full compensation for the nuisance to its users. They also have important implications for the data collection policy which follow below.

Proposition 3 *For a given number of users m (including $m = 1$), we find that*

(i) $\theta_E^* > \theta_N^*$*, that is, the monopolist collects more data with negative externalities compared to the data with no externalities, which is exactly the opposite direction to the social planner's choice* ($\theta_E^s \leq \theta_N^s$) .

(ii) $\theta_E^s < \theta_E^*$ *and* $\theta_N^s \geq \theta_N^*$*, that is, the monopolist collects too much data with negative externalities and too little data with no externalities.*

(iii) $h^* > h^s$*, that is, the monopolist gathers too much data.*

Proof. *The comparison of the two first order conditions with respect to* θ_E^* *and* θ_N^**, (17) and (18), immediately yields that*

$$\theta_E^* > \theta_N^*.$$

We prove that $h^* > h^s$ *by contradiction. Suppose that* $h^s \geq h^*$ *on the contrary. Then, we have* $\theta_N^* \geq \theta_N^s$ *because R is concave in h:* $\theta_N^* > \theta_N^s$ *if R is strictly concave. We know that* $\theta_E^* > \theta_N^*$ *and* $\theta_N^s \geq \theta_E^s$*, which implies that* $\theta_E^* > \theta_N^* \geq \theta_N^s \geq \theta_E^s$*. Then, it must be true that*

$$h^* = \alpha G(\theta_E^*) + (1-\alpha)G(\theta_N^*) > \alpha G(\theta_E^s) + (1-\alpha)G(\theta_N^s) = h^s,$$

which is a contradiction.

Since we have $h^* > h^s$, this in turn implies that $\theta_N^* \leq \theta_N^s$: the inequality is strict of R is strictly concave. Then, for $h^* > h^s$ to hold, it must be true that $\theta_E^* > \theta_E^s$. ■

Proposition 3(i) states that the monopolist collects more types of data with negative externalities relative to data with no externalities, which is exactly the opposite direction to the social planner's choices ($\theta_E^s < \theta_N^s$). Proposition 3(ii) states that the monopolist gathers too much data with information externalities, $\theta_E^s < \theta_E^*$ and too little data with no externalities $\theta_N^* < \theta_N^s$, compared to the socially optimal levels. Finally, Proposition 3(iii) indicates that the overall aggregate amount of data collected by the monopolist is excessive.

Example 1 (continued). Consider again example 1.

The socially optimal outcome θ^s given m satisfies

$$(r+b)m = \eta\left[\theta^s m^2 + \xi\theta^s m(1-m)\right],$$

which leads to

$$\theta^s(m) = \min\left\{1, \frac{(r+b)}{\eta\left[\xi + (1-\xi)m\right]}\right\}, \tag{20}$$

The monopolist's outcome θ^* given m satisfies

$$(r+b)m = \eta(1-\xi)\theta^* m^2,$$

which leads to

$$\theta^*(m) = \min\left\{1, \frac{(r+b)}{\eta(1-\xi)m}\right\} \tag{21}$$

Hence, we have $\theta^s(m) < \theta^*(m)$ for any given m unless $\theta^s(m) = 1$.

Suppose $r+b > 1+\eta(1-\xi)$. Then, from (21), $\theta^*(m) = 1$ for any $m > 0$. This in turn implies, together with (10), $m^* = 1$. So we have

$$m^* = \theta^* = 1.$$

However, even if $r+b > 1+\eta(1-\xi)$, we can have $r+b < \eta(1-\xi/2)$. Then, from (20), we have $\theta^s(m=1) < 1$ and, from (9), $m^s(\theta = 1) < 1$. Therefore, there exist parameters in which

$$\max\left\{m^s(\theta^s), \theta^s(m^s)\right\} < 1.$$

Remark (Positive Information Externalities) We can consider positive (instead of negative) information externalities on non-users. For instance, suppose that part of the benefit

generated by the E category data is shared by non-users as follows:

$$\widehat{B}(\theta_E, m) = \alpha\xi \int_0^{\theta_E} b(\theta, m) dG.$$

Then, it is straightforward that we will have the following relationship between social marginal costs and private marginal costs:

$$SMC_E < SMC_N = PMC_N < PMC_E.$$

As in the case of negative information externalities, we have $PMC_N = SMC_N$. However, the social marginal cost is lower for the E category data than for the N category data as the externalities from the former are positive. In contrast, the private marginal cost is higher for the E category data than for the N category data since the positive externalities from the former increases the reservation utilities of non-users and hence reduces what the monopolist can extract. In summary, as the private marginal cost is weakly larger than the social one for each category, there is no excessive loss of privacy.

6 Data Brokerage Firms and Big Data

In the previous sections, we have considered a monopolistic platform and demonstrated its incentives to collect too much personal data with negative informational externalities. In the Appendix, we show that the same mechanism can be operative even in a market structure with small websites. In particular, we develop a model with a continuum of small websites in order to explore the role of data brokerage firms in the aggregation of information. We show that even if each website alone has no incentives to collect personal data due to its small scale of operation, the emergence of data brokerage markets that purchase and aggregate data from multiple websites can restore incentives to collect personal data and lead to excessive loss of privacy.

The intuition behind this result is the same as in the monopoly model. The information externalities from users on non-users lower each consumer's reservation utility evaluated when her data is not provided. In a model with free entry, we show that entrants can lower compensation to be made for consumer nuisance by the reduction in the reservation utility, even without any business stealing effects as in Mankiw and Whinston (1986). As a result, we can find an equilibrium in which too many websites enter to collect and sell personal data to data brokers.

Taken together, both the monopoly platform model and the data brokerage model with small websites suggest excessive incentives to collect personal information in the presence of information externalities and propose consistent policy remedies, as we discuss in the

next section.

7 Privacy and Data Policy Implications

To deliver policy implications for privacy and data collection from our model in a more effective manner, here we assume that $\underline{u}$ is sufficiently high that the market is covered in both the monopoly outcome and the socially optimal outcome, i.e., we consider the case of $m^s = m^* = 1$.[20] The full market coverage assumption enables us to focus on the information collection aspect of the model and sharpen our results. Also, we now simplify the notation by suppressing the dependence of all relevant functions on m in this section.

Regarding the socially optimal policy, from Proposition 2, we have $\theta_E^s = \theta_N^s = \theta^s$, which means that the social planner will choose the same cutoff for both types of data. This is because there is no externality onto non-users since everyone is a user.

For the monopolist, combining $\theta_E^s = \theta_N^s = \theta^s$ and Proposition 3(ii) leads to

$$\theta_N^* \leq \theta^s < \theta_E^*,$$

where $\theta_N^* < \theta^s$ if R is strictly concave. The non-existence of non-users under full market coverage implies that the socially optimal choice of the data cutoffs depends only on the direct effects of a marginal increase in data collection on the user's benefit and harm. For the data in the N category, there is no externality and thus the reservation utility is also unaffected by a change in the cutoffs. For the data in the E category, by contrast, the positive effect of lowering the reservation utility is taken into account by the monopolist for the choice of θ_E^*, which leads to an excessive data collection policy from the viewpoint of the social planner. The full market coverage assumption yields the following result on the users' net nuisance costs.

Proposition 4 *Suppose that $\underline{u}$ is sufficiently high that the market is covered in both the monopoly outcome and the socially optimal outcome, i.e., we consider the case of $m^s = m^* = 1$. Then, the monopoly generates social inefficiency in that the user's net nuisance costs are larger than the socially optimal level. That is, we have $\Lambda(\theta_E^*, \theta_N^*) > \Lambda(\theta^s, \theta^s)$.*

Proof. *We have*

$$\Lambda(\theta_E^*, \theta_N^*) - \Lambda(\theta^s, \theta^s) = \alpha \int_{\theta^s}^{\theta_E^*} \lambda(\theta) dG - (1-\alpha) \int_{\theta_N^*}^{\theta^s} \lambda(\theta) dG.$$

[20] Alternatively, we can imagine a situation in which the business model of the monopolist is purely ad-financed with zero pricing (or even negative pricing) to ensure all users subscribe.

From Proposition 3(iii), we also know

$$h^* - h^s = \alpha \int_{\theta^s}^{\theta_E^*} \theta dG - (1-\alpha) \int_{\theta_N^*}^{\theta^s} \theta dG > 0.$$

Hence, we have

$$\alpha \int_{\theta^s}^{\theta_E^*} \lambda(\theta) dG > \alpha \int_{\theta^s}^{\theta_E^*} \lambda(\theta^s) dG > (1-\alpha) \int_{\theta_N^*}^{\theta^s} \lambda(\theta^s) dG > (1-\alpha) \int_{\theta_N^*}^{\theta^s} \lambda(\theta) dG$$

which completes the proof of $\Lambda(\theta_E^*, \theta_N^*) > \Lambda(\theta^s, \theta^s)$. ■

We now consider possible regulatory policies to rectify the market inefficiency identified in Proposition 4.

7.1 Opt-in Consent Regulations

To prevent an excessive collection of personal data, "opt-in" regulations under the GDPR require that a customer must actively confirm her consent for data collection with explicit prior permission such as ticking an unchecked opt-in box. Pre-checked boxes that use customer inaction to assume consent are invalid under the GDPR. We assess the effectiveness of such regulations in light of our model. We consider two possible scenarios, depending on whether a firm is allowed to offer a better price to consumers who opt-in.

7.1.1 Opt-in Regulation with Price Discrimination

Consider an opt-in regulation that allows data collection up to the socially optimal level $\theta^s = (\theta_E^s, \theta_N^s) = (\theta^s, \theta^s)$ as a default option. To collect sensitive information beyond that level requires explicit consent by consumers. We first analyze the effectiveness of such policy when the firm is allowed to offer an inducement for the opt-in consent.

With such price discrimination, we show that the regulation can be completely ineffective: there is an equilibrium in which the amount and types of data collected remain the same and the monopolist obtains the same profit as in the absence of such regulation.

To see this, suppose that the monopolist platform provides each user with a price discount of $\delta > 0$ when the user voluntarily agrees to the firm's collection of his E category data in the interval of $[\theta_E^s, \theta_E^*]$, in addition to the basic plan which collects data up to θ_E^s in the E category and up to θ_N^* in the N category. If the user decides to opt-in, his total net nuisance cost is given by

$$\Lambda^{\text{opt-in}} \equiv \Lambda(\theta_E^*, \theta_N^*) = \alpha \int_0^{\theta_E^*} \lambda(\theta) dG + (1-\alpha) \int_0^{\theta_N^*} \lambda(\theta) dG, \tag{22}$$

Alternatively, when the user opts out while all the others opt in, his nuisance cost from

subscribing to the 'basic plan' amounts to

$$\Lambda^{\text{opt-out}} = \alpha \left[\int_0^{\theta^s} \lambda(\theta) dG + \xi \int_{\theta^s}^{\theta_E^*} \psi(\theta) dG \right] + (1-\alpha) \int_0^{\theta_N^*} \lambda(\theta) dG. \tag{23}$$

The term $\xi \int_{\theta^s}^{\theta_E^*} \lambda(\theta) dG$ in the square bracket for $\Lambda^{\text{opt-out}}$ represents the externalities part from other users who opt in. We thus have an equilibrium in which every consumer chooses to opt-in if the following inequality is satisfied:

$$\delta + \alpha \int_{\theta^s}^{\theta_E^*} b(\theta) dG \geq \alpha(1-\xi) \int_{\theta^s}^{\theta_E^*} \psi(\theta) dG \tag{24}$$

The LHS of (24) represents gains from the price reduction and a potentially better service due to the opt-in choice. The RHS represents the higher nuisance cost from the opt-in. Without regulation, the monopolist would choose the data collection policy of $\theta^* = (\theta_E^*, \theta_N^*)$ and charge the price of

$$p = \underline{u} + B(\theta_E^*, \theta_N^*) - \left[\Psi(\theta_E^*, \theta_N^*) - \widehat{\Psi}(\theta_E^*, \theta_N^*) \right] = u - \Lambda(\theta_E^*, \theta_N^*) + \widehat{\Psi}(\theta_E^*, \theta_N^*)$$

Under the opt-in regulation, the monopolist can replicate this outcome with the price of $\widetilde{p} = p + \delta$ along with a discount option of δ for the users who consent to the collection of E category private information belonging to the interval of $[\theta_E^s, \theta_E^*]$, where

$$\delta = \alpha \left[(1-\xi) \int_{\theta^s}^{\theta_E^*} \psi(\theta) dG - \int_{\theta^s}^{\theta_E^*} b(\theta) dG \right]$$

Our analysis thus suggests that opt-in regulations can be sabotaged if the monopolist is allowed to offer a discount to the users who opt in.

7.1.2 Opt-in Regulation without Price Discrimination

The preceding discussion motivates us to consider an alternative regulatory restriction that the platform is constrained to offer the *same* price regardless of the consumer's opt-in choices. As in the case of price discrimination, consider a regulatory policy that requires a collection of data beyond the socially optimal level of $\theta^s = (\theta^s, \theta^s)$ to obtain explicit consent by consumers. Suppose that the monopolist proposes to collect additional E category data in the interval set of $[\theta^s, \widetilde{\theta}_E]$ with opt-in consent where $\widetilde{\theta}_E > \theta^s$. When the monopolist is not allowed to provide a discount to opt-in consumers, we can consider two cases.

If $b(\theta^s) < (1-\xi)\psi(\theta^s)$, it can be easily verified that it is a dominant strategy for every consumer not to opt-in for any $\widetilde{\theta}_E$. This is because even if all other consumers opt in, a consumer finds it optimal not to opt-in under this condition. As a result, opt-in regulations

that require no price discrimination can lead to the socially optimal outcome because the monopolist is able to collect data only up to the socially optimal level.

However, if ξ is sufficiently large and $b(\theta^s) > (1-\xi)\psi(\theta^s)$, the regulation can be less effective due to coordination failure on the users' side. More specifically, the monopoly platform can always find a $\widetilde{\theta}_E$ close enough to θ^s such that it satisfies the following inequality:

$$\int_{\theta^s}^{\widetilde{\theta}_E} b(\theta)dG \geq (1-\xi)\int_{\theta^s}^{\widetilde{\theta}_E} \psi(\theta)dG. \tag{25}$$

Let $\widetilde{\theta}_E^*$ be the the maximum $\widetilde{\theta}_E$ that satisfies the inequality (25). If $\widetilde{\theta}_E^* \geq \theta_E^*$, the monopolist can implement its preferred data policy of $\theta^* = (\theta_E^*, \theta_N^*)$ even under opt-in regulations without price discrimination. If $\widetilde{\theta}_E^* < \theta_E^*$, then the monopolist is able to collect data up to the level of $\widetilde{\theta}_E^*$ with users' explicit consent, which is an improvement over no regulation, but does not achieve the socially optimal outcome.

Our analysis thus suggests that for the information type that has strong information externalities (i.e., high ξ), but is not essential to the service provision (i.e., low $b(.)$), the outright ban may be necessary as an effective policy. On the other hand, for the information type without strong externalities and with modest essentiality, opt-in regulation without price discrimination rule would be sufficient as an alternative (and less paternalistic) policy remedy.

7.2 Informed Consent Approach and EU GDPR

Our model has important policy implications for the ongoing policy debate regarding privacy protection from the collection of personal data by on-line platforms and websites and their sales to third parties such as data brokerage firms. Currently, most countries' privacy regulation and law are based on 'informed consent' approach. This approach finds its justification on the premise that an individual's informed consent provides legitimacy for any information collection and its use. Despite its intuitive appeal, there has been wide criticism against such approach. One argument is that privacy notices are rarely read, and even if read, not easy to fully understand (The White House, 2014).[21] This criticism has naturally led to the discussion of how we can improve transparency about firms' data practices (Federal Trade Commission, 2012). No one shall dispute the importance of improving readability and transparency of data policies. However, our model shows that such approach alone may not warrant an effective enhancement of privacy protection.[22] In our model even

[21] McDonald and Cranor (2008) estimated the total time opportunity cost being worth of $781 billion per year if all web visitors had read all privacy policies.

[22] Solove (2013) points out the ineffectiveness in addressing the current privacy concerns by means of providing consumers with more transparency on their personal data collection and use.

costless reading and *perfect* understanding lead to an equilibrium with an excessive privacy loss in the presence of strong negative information externalities.

Consistent with our analysis, the recent global policy trend appears to improve on the current informed consent approach. The most prominent example is the EU General Data Protection Regulation (GDPR), which imposes strict new rules on controlling and processing personally identifiable information to protect personal data and privacy for all individuals within EU. According to the GDPR Article 4, 'consent' of the data subject requires "*freely given, specific, informed*, and *unambiguous* indication ... by a statement or by a clear *affirmative* action (italics added)." In particular, Article 7 specifies that for consent to be considered *freely given*, data controllers should not withhold or offer a degraded version of service for data subjects who refuse or later withdraw consent, except the personal data that is essential of the provision of a service. This specification is in accordance with our analysis that shows the ineffectiveness of opt-in regulations with price discrimination: the platform may collect data that is essential for the provision of a service, but there should be no monetary inducement for opting-in (or equivalently, no penalty for not opting-in). However, our analysis also suggests that such explicit consent regulation may fall short in limiting the collection of personal data only up to the socially optimal level in the presence of strong information externalities.

Our data brokerage model also suggests that banning data trade may be a remedy, in particular, for the type of data with strong externalities. In the current GDPR, processing of data by a third party is allowed only "for the purposes of the legitimate interests pursued ... by a third party" (GDPR Article 6.1(f)) in the absence of the data subject's explicit consent. Most experts of GDPR, however, interpret that the contextualization for applying the relief via the legitimate interests will be extremely limited and thus most cases will require the informed consent by data subjects. Thus, the new GDPR regulation makes the data processing by marketing and sales organizations much costly, which can mitigate the problem of an excessive entry of small websites who use data monetization as their business model.

8 Concluding Remarks

As our lifestyle becomes increasingly reliant on the Internet, our daily activities through all kinds of computer and mobile devices leave digital trails, constantly producing up-to-date information about our activities. Such data becomes so valuable that now many websites and content providers offer their content for free or at a highly subsidized price in exchange for users' agreement to more or less uncommitted use of personal data, and the collected data is handed over to the data broker markets. This has raised critical privacy

concerns about potential harms and costs to individuals and society.

In this paper we provide a model of privacy based on the concept of information externalities. Even if data collection requires consumers' consent and consumers are fully aware of the consequences of such consent, we show that the market equilibrium is characterized by an excessive collection of personal information and the resulting loss of privacy compared to the social optimum. Therefore, we find that the current main privacy regulatory framework of the informed consent model may be ineffective to address the privacy concerns associated with the data broker industry.

To quote Schneier (p.238), "[d]ata is the pollution problem of the information age, and protecting privacy is the environmental challenge." As the pollution problem of the industrial age challenges us economists to come up with various policies—either market-oriented mechanisms or direct regulations—we now need to take a similar approach to the personal data. As pollutants have negative externalities and any preventive efforts such as abatement have the public good problem, the privacy protection in this big data world generates information externalities and the privacy protection may be viewed as a public good. We hope that our research provides a step conducive to more research in this direction.

References

Acquisti, A. (2009). Nudging privacy: The behavioral economics of personal information. *IEEE security & privacy 7*(6).

Acquisti, A. and J. Grossklags (2004). Privacy attitudes and privacy behavior. In *Economics of information security*, pp. 165–178. Springer.

Acquisti, A. and J. Grossklags (2007). What can behavioral economics teach us about privacy. *Digital Privacy: Theory, Technologies and Practices 18*, 363–377.

Acquisti, A., C. R. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 52*(2), 442–92.

Athey, S. (2014). Information, privacy and the internet: an economic perspective. Central Planning Bureau (CPB) Netherlands Bureau for Economic Policy Analysis.

Bataineh, A. S., R. Mizouni, M. El Barachi, and J. Bentahar (2016). Monetizing personal data: A two-sided market approach. *Procedia Computer Science 83*, 472–479.

Bergemann, D. and A. Bonatti (2015). Selling cookies. *American Economic Journal: Microeconomics 7*(3), 259–294.

Bloch, F. and G. Demange (2018). Taxation and privacy protection on internet platforms. *Journal of Public Economic Theory 20*(1), 52–66.

Bourreau, M., B. Caillaud, and R. Nijs (2018, February). Taxation of a digital monopoly platform. *Journal of Public Economic Theory 20*(1), 40–51.

Campbell, J., A. Goldfarb, and C. Tucker (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy 24*(1), 47–73.

Casadesus-Masanell, R. and A. Hervas-Drane (2015). Competing with privacy. *Management Science 61*(1), 229–246.

Council of Economic Advisers (2015). Big data and differential pricing.

Daughety, A. F. and J. F. Reinganum (2010). Public goods, social pressure, and the choice between privacy and publicity. *American Economic Journal: Microeconomics 2*(2), 191–221.

DellaVigna, S. and U. Malmendier (2004). Contract design and self-control: Theory and evidence. *The Quarterly Journal of Economics 119*(2), 353–402.

Erlich, Y., T. Shor, I. Pe'er, and S. Carmi (2018). Identity inference of genomic data using long-range familial searches. *Science*.

Fairfield, J. A. and C. Engel (2015). Privacy as a public good. *Duke Law Journal 65*, 385–457. Available at: https://scholarship.law.duke.edu/dlj/vol65/iss3/1.

Goh, K.-Y., K.-L. Hui, and I. P. Png (2015). Privacy and marketing externalities: evidence from do not call. *Management Science 61*(12), 2982–3000.

Hirsch, D. D. (2006). Protecting the inner environment: What privacy regulation can learn from environmental law. *Georgia Law Review 41*(1). https://ssrn.com/abstract=1021623.

Jernigan, C. and B. F. Mistree (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday 14*(10).

Johnson, J. P. (2013). Targeted advertising and advertising avoidance. *The RAND Journal of Economics 44*(1), 128–144.

Kosinski, M., D. Stillwell, and T. Graepel (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences 110*(15), 5802–5805.

Lane, J., V. Stodden, S. Bender, and H. Nissenbaum (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press.

Lefouili, Y. and Y. L. Toh (2017). Privacy and quality. *TSE Working Paper 17*(795).

MacCarthy, M. (2011). New directions in privacy: Disclosure, unfairness and externalities. *I/S: A Journal of Law and Policy for the Information Society 6*, 425–512.

Mankiw, N. G. and M. D. Whinston (1986). Free entry and social inefficiency. *The RAND Journal of Economics 17*(1), 48–58.

McDonald, A. M. and L. F. Cranor (2008). The cost of reading privacy policies. *ISJLP 4*, 543.

Montes, R., W. Sand-Zantman, and T. M. Valletti (2015). The value of personal information in markets with endogenous privacy. *Management Science (forthcoming)*.

Posner, R. A. (1978). The right of privacy. *Georgia Law Review 12*(3), 393–422.

Posner, R. A. (1981). The economics of privacy. *American Economic Review 81*(2), 405–409.

Preibusch, S., K. Krol, and A. R. Beresford (2013). The privacy economics of voluntary over-disclosure in web forms. In *The Economics of Information Security and Privacy*, pp. 183–209. Springer.

Reimers, I. and B. Shiller (2018). Proprietary data, competition, and consumer effort: An application to telematics in auto insurance.

Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company.

Singer, E., J. Van Hoewyk, R. Tourangeau, D. M. Steiger, M. Montgomery, and R. Montgomery (2001). Final report on the 1999-2000 surveys of privacy attitudes.

Smith, H. J., T. Dinev, and H. Xu (2011). Information privacy research: an interdisciplinary review. *MIS quarterly 35*(4), 989–1016.

Solove, D. J. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review 126*, 1879–2479.

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are.* HarperCollins. May.

Stigler, G. J. (1980). An introduction to privacy in economics and politics. *Journal of Legal Studies 9*(4), 623–644.

The White House (2014). Big data: Seizing opportunities, preserving values.

Thisse, J.-F. and X. Vives (1988). On the strategic choice of spatial price policy. *The American Economic Review*, 122–137.

Tucker, C. (2017). Privacy, algorithms and artificial intelligence. In *Economics of Artificial Intelligence*. University of Chicago Press.

US Senate (2013). A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes.

Waldo, J., H. Lin, and L. I. Millett (2007). *Engaging privacy and information technology in a digital age*. National Academies Press.

Zarsky, T. Z. (2018). An analytic challenge: Discrimination theory in the age of predictive analytics. *I/S: A Journal of Law and Policy for the Information Society 14*, 11.

Appendix: The Analysis for Data Brokerage with Many Small Websites

The Model

Consider a mass one of small but monopolistic websites in their own niche market and a mass one of consumers. All consumers are homogeneous, and they may patronize multiple websites. The websites are heterogeneous regarding the value their content generates for consumers: let v denote a consumer's valuation of content, which is assumed to be distributed according to a distribution function F with density f over the interval $[\underline{v}, \overline{v}]$ with $\underline{v} > 0$. We assume that all websites have the same fixed cost of entry $K > 0$. We envision a situation in which the types of information collected across websites are different. Suppose that all websites use the business model of data monetization. Then, let $R(n)$ denote the aggregate revenue of the data brokers, where n is the measure of websites who feed the personal information about their users to data brokers.[23]

[23]We make distinctions between websites that collect information for sale to the data brokerage firm and the firms that would purchase consumer specific data from the brokerage firm. Our focus is on the information collection side, with potential competition among data purchasers and how the data brokerage firm sells data to them in the black box. All we need for our model would be that the data brokerage firm's revenue $R(n)$ will increase with more data and the marginal value of additional data collection is decreasing in the relevant range.

As in our analysis in the main text, we assume that there are two categories of websites depending on the nature of information collected. More specifically, for each type of website generating value v, there are α fraction of websites (E category websites) that collect information with externalities, whereas $(1-\alpha)$ fraction of websites (N category websites) collect information without externalities. All websites in the same category are assumed to generate the same level of nuisance to the users. More precisely, let S_E and S_N denote the sets of entrants in the E and N categories (with the corresponding measures of n_E and n_N), respectively. Let $\Phi(x_E, x_N; n_E)$ represent the nuisance cost inflicted on consumer i when the consumer uses a measure (x_E, x_N) (with $x_E \in [0, n_E]$ and $x_N \in [0, n_N]$) from each category of websites given that all other consumers patronize all websites in the E category of measure n_E. We assume that the total nuisance cost is additive (as in (2) for the main model), that is, it is the sum of the nuisance from each category of data (i.e., website).

$$\Phi(x_E, x_N; n_E) = \phi_N(x_N) + \phi_E(x_E; n_E),$$

where ϕ_c represents the nuisance cost from the c category of data for $c \in \{E, N\}$. Note that the nuisance from the N category data depends only on x_N whereas the nuisance from the E category data depends on both x_E and n_E due to information externalities. We further assume the following:

A1: $\phi_E(x_E; n_E)$ strictly increases in each element and concave in x_E with $\phi_E(0,0) = 0$.

A2: $\phi_N(x) = \phi_E(x,x) = \kappa x$ and $\phi_E(0, n_E) = \xi n_E$ with $\kappa > \xi > 0$.

A2 states that as long as the consumer uses all the websites, the types of data do not matter. The linearity of $\phi_N(x)$ and $\phi_E(0, n_E)$ is assumed for the consistency with the additive structure in the nuisance cost. Note that $\phi_E(0, n_E)$ captures the information externalities on a non-user from users. As an example satisfying A1 and A2, we can consider the following CES nuisance cost of

$$\phi_E(x, n) = \kappa[\beta x^\rho + (1-\beta) n^\rho]^{\frac{1}{\rho}},$$

where $0 < \beta < 1$ and $\rho < 1$. With this specification, it can be easily verified that $\phi_E(n, n) = \kappa n$. Thus, we have $\Phi(n_E, n_N; n_E) = \kappa \cdot (n_E + n_N)$. The linearity of $\phi(0, n_E)$ is satisfied with $\xi = (1-\beta)^{\frac{1}{\rho}} \kappa$.

In addition, we assume scale economies in data brokerage.

A3: $R'(0) < \frac{\partial \phi_E(0,0)}{\partial x_E}$ and $R(.)$ is strictly concave in the relevant range.

A3 states that in the absence of the brokerage industry, no website (of measure zero) has incentives to collect data on its own (alternatively, in the presence of the brokerage industry if no other website sells data for aggregation), because the size of the data that can be collected and monetized by each firm is too small to justify the nuisance costs.

This implies that each website should adopt the pure content pricing model without data collection and processing. This starting point is purposefully set this way because we aim to show the role of data brokerage firms to revive each firm's incentives for data monetization.

We consider a three-stage game with the following timing. In Stage 1, each website simultaneously decides whether to be active or not: to be active, a website must incur the fixed cost of entry $K > 0$. In Stage 2, each active website simultaneously chooses its business model (from either data monetization or pure content pricing) and the content price. In Stage 3, each consumer decides which websites to patronize among the active websites given each firm's privacy policy and content price offers.

Competitive Personal Data Monetization

Let us first analyze the case that all active websites in the set S_E and S_N of measures n_E and n_N adopt the business model of data monetization at Stage 2. We focus on characterizing the equilibrium in which every consumer patronizes all active websites. Let i and i' represent two different websites in S_E with v^i and $v^{i'}$. Let p^i and $p^{i'}$ denote the respective content prices they charge. Since all websites are identical in terms of the nuisance cost they generate from data sales, for any pair of websites (i,i') in equilibrium we have

$$v_E(n_E) := v^i - p_E^i = v^{i'} - p_E^{i'}. \tag{26}$$

A similar logic applies to any pair of websites j and j' in S_N.

$$v_N(n_N) := v^j - p_N^j = v^{j'} - p_N^{j'}. \tag{27}$$

Hence, each consumer's payoff in equilibrium will be given by

$$\begin{aligned} & \int_{i\in S_E} \left(v^i - p_E^i\right) di + \int_{j\in S_N} \left(v^j - p_N^j\right) dj - \Phi(n_E, n_N; n_E) \\ = \;& n_E \cdot v_E(n_E) + n_N \cdot v_N(n_N) - \kappa \cdot (n_E + n_N). \end{aligned}$$

Consider now Stage 3 in which the vector of prices generated by the system (26)-(27) are given from Stage 2. Define the utility that a consumer obtains from choosing $x_E \in [0, n_E]$ and $x_N \in [0, n_N]$ websites when all other consumers patronize all websites in the E category of measure n_E:

$$u(x_E, x_N; n_E) = x_E \cdot v_E(n_E) + x_N \cdot v_N(n_N) - \kappa x_N - \phi_E(x_E; n_E).$$

As we assumed in A1-A3, $u(x_E, x_N; n_E)$ is linear in x_N and is convex in x_E.

For websites in the N category, the consumers' patronage decisions are independent of

others' because there are no information externalities. As long as $v(n_N) \geq \kappa$, all consumers patronize all n_N websites in the N category. Each website in S_N faces no competition and generates no externalities. Hence, at Stage 2, each will charge the highest price that makes a consumer indifferent, meaning $v_N(n_N) = \kappa$.

For websites in the E category, all consumers patronize all n_E websites if the following incentive constraint is satisfied for all $x_E \in [0, n_E]$:

$$[IC : (x_E; n_E)] \quad u(n_E, n_N; n_E) \geq u(x_E, n_N; n_E). \tag{28}$$

The convexity of $u(x_E, n_N; n_E)$ with respect to x_E means that $u(x_E, n_N; n_E)$ is maximized either at $x_E = 0$ or at $x_E = n_E$. Hence, the IC condition will be satisfied for any $x_E \in [0, n_E]$ if $[IC : (0; n_E)]$ is satisfied, which is given by

$$n_E \cdot v_E(n_E) - \kappa n_E \geq -\phi_E(0, n_E). \tag{29}$$

In equilibrium, inequality (29) must hold with equality because otherwise each website finds an incentive to raise its price at Stage 2. As the IC condition is binding, the necessary condition (28) implies

$$v_E(n_E) = \frac{\phi_E(n_E, n_E) - \phi_E(0, n_E)}{n_E} (= \kappa - \xi) > 0. \tag{30}$$

In summary, we derived the condition that determines the surplus from each website in the proposed equilibrium depending on the type of website. Any website i in S_E charges a price equal to

$$\hat{p}_E^i = v^i - (\kappa - \xi) \text{ for } \forall i.$$

For any website j in S_N, the equilibrium price for the content service is derived as

$$\hat{p}_N^j = v^j - \kappa \text{ for } \forall j. \tag{31}$$

In addition, the rent that each website will receive from selling its data to the brokerage market is equal to $R'(n_E + n_N)$, the marginal contribution of its data to the data aggregation for a competitive brokerage market. Then, each website i's equilibrium payoff is equal to $R'(n_E + n_N) + \hat{p}_E^i$ or $R'(n_E + n_N) + \hat{p}_N^i$.

As the last step, let us check the deviation incentive by website i at Stage 2 to a pure content pricing model. Obviously, websites in S_N have no incentives to deviate as long as $R'(n_E + n_N) > \kappa$. Consider a potential deviation of website $i \in S_E$ to the pure content pricing model. When the pure content pricing model is adopted by the website and hence

no data is collected to sell to the brokerage firm, each consumer's willingness to pay for the website is v^i. This implies that the deviation is not profitable for $i \in S_E$ if

$$R'(n_E + n_N) + \hat{p}_E^i \geq v^i \Longleftrightarrow R'(n_E + n_N) \geq \kappa - \xi.$$

Therefore, as long as no website in S_N has any incentive to deviate, neither does any website in S_E.

Excessive Entry of Websites

Our analysis so far has been confined to the ex post entry stage (from Stage 2) when a fixed measure of websites entered the market at Stage 1. Let us move backwards and study Stage 1 by making the entry decision endogenous. Let (n_E^*, n_N^*) be the equilibrium measure of websites. Then, the marginal website's value to consumers in the E category, v_E^*, is given by $\alpha[1 - F(v_E^*)] = n_E^*$; similarly, we have $(1-\alpha)[1 - F(v_N^*)] = n_N^*$ in the N category. This implies that in the first stage, the extent of entry is determined by the following two free-entry conditions:

$$v_E^* - \kappa - \xi + R'(n^*) = K, \tag{32}$$

$$v_N^* - \kappa + R'(n^*) = K, \tag{33}$$

where K is the fixed cost of entry and $n^* = n_E^* + n_N^* = \alpha[1 - F(v_E^*)] + (1-\alpha)[1 - F(v_N^*)]$. We have $v_N^* > v_E^*$: the threshold value is lower for the E category websites than that for the N category.

Given the cutoff types of entrant in each category, v_E and v_N, social welfare can be written as follows.

$$W(v_E, v_N) = \alpha \int_{v_E}^{\overline{v}} x dF(x) + (1-\alpha) \int_{v_N}^{\overline{v}} x dF(x) + R(n) - (\kappa + K) n$$

where $n = \alpha[1 - F(v_E)] + (1-\alpha)[1 - F(v_N)]$ measures the total number of active websites and κ is nuisance cost per website. The welfare-maximizing cutoff types (v_E^s, v_N^s) can be derived by the following first-order conditions:.

$$-v_i f(v_i^s) - R'(n) f(v_i^s) + (\kappa + K) f(v_i^s) = 0, \tag{34}$$

where $i = E, N$. Hence, $v_E^s = v_N^s = v^s$ is given by

$$v^s + R'(1 - F(v^s)) = \kappa + K. \tag{35}$$

Let $1-F(v^s) \equiv n^s$. The comparison of (32)-(33)and (35) reveals

$$n^s < n^* \text{ and } v_E^* < v^s < v_N^*.$$

Under A1–A3, there is an excessive entry of websites for the E category, but an insufficient entry for the N category (i.e., $v_E^* < v^s < v_N^*$). Overall, however, there is an excessive entry in that $n^* > n^s$. This result can be proved by contradiction. Suppose that $n^s \geq n^*$. Then, from the strict concavity of $R(.)$ (see A3), we have both $v_N^* \leq v^s$ and $v_E^* < v^s$, which means $n^s < n_E^* + n_N^* = n^*$, a contradiction. We thus have $n^* > n^s$. The concavity of $R(.)$ then implies, from (32) and (34), that $v_N^* > v^s$. Then, to satisfy $n^* > n^s$, $v^s > v_E^*$ must hold.